Expanding a Mexican-Food restaurant franchise.

Introduction

A group of investors is considering to expand the number of restaurants in new locations over the world. The restaurant brand is ranked 20 in the top 50 best Mexican-food restaurants. The restaurants offer luxury commodities and amenities, as well as a great variety of international food and well-know chefs of around the world. The target customer is focused on people with high-spending-profiles living in big cities.

As such, investors need information on potential new locations to open new restaurants. These potential locations need to meet some basic criteria:

1. Locations should be located in cities with more than 1 million inhabitants.
2. The average income of the population should exceed $50,000.00 dollars per year.
3. Cities should be considered multi-cultural.
4. Other competitors should be established in the surroundings areas.

In this respect, investors are interested in the city of Toronto, Canada. Toronto is a multi-cultural city whose residents hold a higher income-per-family with respect to the rest of cities in Canada1. In addition, the culinary offer in Toronto is vast, ranging from Japanesse to Babarian food restaurants and from cheap to very expensive venues.

In order to fulfil the information requirements by the investors, the data science group will apply some clustering techniques to classify neighbourhoods given venues and demographic data of the city of Toronto.

Data Analysis

Data provenance

The group will be looking at different data sources, insofar a single dataset with all the required information may not be public available. The datasets used are provided as open data from Canada government aggencies or, obtained by scrapping some web sites. We will focusing on geo-location data, zip codes and demographic (i.e., income) of Toronto. In particular, we will be looking at the following datasets:

Toronto zip codes. We scrapped a Wikipedia page containing data of zip codes of Toronto, Canada. (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). This dataset contains the following attributes:

- Postal Code
- Borough
- Neighbourhood

Toronto geo-data. We used a provided dataset containing geo-localization data for every postal code in Toronto. This dataset is as follows:

- Postal Code
- Latitude
- Longitude

Ward geo-data. We used an open dataset containing information on the ward in the city of Toronto (https://open.toronto.ca/dataset/ward-profiles-2018-25-ward-model/). This dataset includes geo-localization coordinates (i.e., polygon) for each ward in Toronto. The schema is as follows:

- Ward_number
- Area
- Geometry

Income data. We used an open data dataset containing demographic information of Toronto (https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/ward-profiles/). In particular, it contains the average income-per-family for every ward in the city of Toronto. The schema is as follows:

- Ward_number
- Average_income

Venue data. We used Foursquare api to obtain data of venues in Toronto. In a nutshell, we obtain data on the venue location, categories and type of venue.

Wrangling process

As I mentioned before, we are using data from different heterogenous datasources. Thus, I wrangled data to obtain a uniform view of the target schema. We need data that associates:

- Postal code, Wards and Neighbourhoods of Toronto.
- Venues categories.
- Venues locations.
- Income-per-family in wards and areas of Toronto.

General process:

First, we obtain the geo-coordinates for each postal code by joining the postal code and geo-data datasources.

Second, given the coordinates for each postal code we determine the ward to which the neigbourhood belongs. Here, we use the ward datasource that contains a polygon that represents the ward area. For each postal code coordinate, we calculate the intersection with each polygon.

Third, we obtain the average income-per-family of each ward by joining the ward with the ward-income datasource. Here, we discretized the income variable into three bins, representing low, medium and high income.

Fourth, we obtain venue information using Foursquare for each geo-coordinate representing a Neighbourhood.

Finally, we merge all data resulting in a dataset containing, venue information, location and average income.

Clustering data

In order to obtain a view of the potential locations in which the new restaurants can be open, we run a clustering algorithm to group similar venues given its location, categories and the average income of the population in such areas. In particular, we use k-means algorithm. We parametrized the clustering algorithm using two values of “k”, namely, 3 and 4.

Results

By running the algorithm using different values of k, we noticed that any value greater than three resulted in 3 distinct clusters. Interestingly, this number of clusters also corresponds to the number of bins on the income variable.

Moreover, we can see that most of the international-food venues, are concentrated in a small area, namely, in the center of the city, which also happens to be one of the areas with the highest income-per-familiy.

By analysing this data, investors decided to open a new restaurant in the area of Toronto Dominion Centre, because there are several international restaurants surrounded by wards with high income.